Asymptotic behaviors of the loss probability for a finite buffer queue with QBD structure

Masakiyo Miyazawa, Yutaka Sakuma and Syuhei Yamaguchi
Department of Information Sciences
Tokyo University of Science

Abstract

We are concerned with a queue with a finite buffer and arrival and service processes that are generated by a continuous time Markov chain with finitely many states. This model includes a many server queue, and is described by a truncated QBD, where QBD stands for a quasi-birth-and-death process. It is shown that the loss probability of this queue geometrically decays with a constant prefactor as the buffer size goes to infinity if the corresponding queue with an unlimited buffer is stable. For this, we firstly consider asymptotic behaviors of the truncated QBD. We then specialize this result for a many server queue with a finite buffer and Markovian arrival and service processes.

1. Introduction

Loss probabilities are important for performance evaluations of queueing systems with finite buffers, where buffers only accommodate waiting customers. In principle, the loss probabilities can be computed numerically since the state spaces are essentially finite, but closed form formulas are hard to get for them except for special cases such as $M/M/c/K+c$ and $M/PH/1/K+1$ queues, where K's are buffer sizes. So, heuristic approximations have been proposed for them (e.g., see [5, 13, 17]). However, the loss probabilities become very small for the case of large buffers, so it is interesting to see their exact asymptotic behaviors. There are some studies in this direction. In particular, Baiocchi [4] showed that the loss probability geometrically decays with a constant prefactor in the $MArP/G/1$ queue under a light-tail condition on the service time distribution, where $MArP$ stands for a Markovian arrival process, and will be explained below. More specific results are obtained for the $M/G/1/K+1$ and $GI/M/1/K+1$ queues in [3] (see also [6]).

In this paper, we study this asymptotics of the loss probability from a different perspective. We assume that the queue has a single buffer of finite size, but do not assume a single server. Service discipline is not essential as long as certain requirements are satisfied, but we suppose First-Come First-Served in mind. It is noticed that the single server assumption plays crucial roles in the existence results cited above (see [3, 4, 6]). There are two reasons for this. First, it allows to describe the models by embedded Markov chains at arrival or departure instants, which simplifies the state spaces. Secondly, the

loss probabilities are easily computed from the stationary probabilities that the systems are empty. We can not use these properties for many server queues in general.

So far, we attack the asymptotic problem, using a different approach and assuming slightly simpler assumptions on arrivals and services. Namely, we assume that arrival times of customers and their service completion times are driven by a continuous time Markov chain with finitely many background states. In the queueing literature, this class of arrival processes are called Markovian arrival processes, and MAP has been used for short in the literature. In this paper, we refer to it as MArP to distinguish from a Markov additive process, which plays key roles in this paper. The corresponding process for services is referred to as a Markovian service process, MSP for short. Under this situation, we can describe the finite queue by a truncated quasi-birth-and-death process, a truncated QBD for short. So, we first consider asymptotic behaviors of the stationary probabilities at the truncated level in a truncated QBD as the truncated level goes to infinity. This may have an independent interest.

Our major tools are occupation measures before hitting given sets of states for the continuous time Markov additive process that generates the truncated QBD. These measures are relatively easy to handle for studying the asymptotic behaviors since they are monotone as the hitting states going away. It is already known that the stationary distribution of the truncated QBD can be obtained in terms of the occupation measures. Using this result, we show that the stationary probabilities at the truncated level geometrically decay with constant prefactors, provided that the corresponding QBD, whose level is unlimited, is stable, that is, it has the stationary distribution. Some remarks are also given for the unstable case. We then apply this result to a c server queue with a buffer of size K, Markovian arrival and Markovian service processes, which is denoted by $MArP/MSP/c/K+c$. In this way, we generalize Theorem 1 of [4] for many server queues.

The remainder of this paper is composed of four sections. In Section 2, we introduce a Markov additive process with finitely many background states and the QBD generated by this additive process. We also prepare some asymptotic results for them. In Section 3, we truncate the QBD, and obtain its asymptotics as the truncation level goes to infinite. In Section 4, we consider the loss probability for $MArP/MSP/c/K+c$, and obtain its exact asymptotics as the buffer size becomes large. In Section 5, we give some remarks on possible extensions of the present results.

2. Markov additive and QBD processes

We first give our vector and matrix notation since they are frequently used. For vector $\boldsymbol{x}$, its i-the entry is denoted by $\boldsymbol{x}(i)$. Similarly, $A(i,j)$ denotes the ij-th entry of a matrix A. All inequalities of vectors and matrices are component wise. For example, $\boldsymbol{x} \geq \mathbf{0}$ means that $\boldsymbol{x}(i) \geq 0$ for every entry i. A row vector $\boldsymbol{x}$ is called a probability vector or distribution if it is nonnegative and $\boldsymbol{x}\mathbf{1} \leq 1$, where $\mathbf{1}$ is the column vector of an appropriate size whose entries are all units. A nonnegative matrix A is called a transition probability matrix if $A\mathbf{1} \leq \mathbf{1}$. If this inequality is strict for some entries, then A is called defective. A square matrix A is called a transition rate matrix if $A\mathbf{1} \leq \mathbf{0}$ and all off-diagonal entries of A are

nonnegative. If the inequality $A\mathbf{1} \leq \mathbf{0}$ is strict in some entries, then A is called defective.

Let $\mathcal{N}$ be the set of all integers and let $\mathcal{S}_B$ be a finite set. We define a continuous time process $\{(X(t), B(t))\}_{t\geq 0}$ with state space $\mathcal{N} \times \mathcal{S}_B$ in the following way. Assume that $B(t)$ is a continuous time Markov chain. We decompose the transition rate matrix of this Markov chain into three $\mathcal{S}_B \times \mathcal{S}_B$ matrices, Q_{-1}, Q_0 and Q_1 in such a way that Q_1 and Q_{-1} are nonnegative and Q_0 is a defective transition rate matrix. Thus, $B(t)$ has the transition rate matrix $Q_{-1} + Q_0 + Q_1$. Throughout the paper, it is assumed that $Q_{-1} + Q_0 + Q_1$ is non-defective and irreducible. The integer valued process $X(t)$ increases by i when $B(t)$ changes according to Q_i, respectively, for $i = 0, \pm 1$. The process $(X(t), B(t))$ is said to be a Markov additive process. A definition for a general Markov additive process can be found in [7], but descriptions there are too abstract for our arguments. Note that $(X(t), B(t))$ is also a Markov chain. In this paper, we refer to $X(t)$ and $B(t)$ as level and background processes, respectively.

We next introduce matrices for occupation measures for the Markov additive process $(X(t), B(t))$. For $A \subset \mathcal{N}$, let $\tau_A^c = \inf\{t > 0; X(t) \notin A\}$. Define matrix $N_{k\ell}^A$ as

$$N_{k\ell}^A(i,j) = E_{(k,i)}\left[\int_0^{\tau_A^c} \boldsymbol{I}(X(t) = \ell, B(t) = j)dt\right], \qquad k, \ell \in A,$$

for $i, j \in \mathcal{S}_B$, where $E_{(k,i)}$ stands for the conditional expectation given that $X(0) = k, B(0) = i$, and $\boldsymbol{I}(\cdot)$ is the indicator function of statement "$\cdot$". For given starting state (k, i), $\{N_{k\ell}^A(i,j); \ell \geq 1, j \in \mathcal{S}_B\}$ is known as an occupation measure before hitting $A^c \equiv \mathcal{N} \setminus A$ (e.g., see [1]). Let

$$(0,+) = \{n \in \mathcal{N}; n \geq 1\}, \qquad (-,m) = \{n \in \mathcal{N}; n \leq m-1\},$$
$$(0,m) = \{n \in \mathcal{N}; 1 \leq n \leq m-1\}.$$

We will use these sets for the A. Let

$$R = Q_1 N_{11}^{(0,+)}.$$

It is well known that the matrix R is obtained as the minimal nonnegative solution for the matrix equation:

$$Q_1 + RQ_0 + R^2 Q_{-1} = O. \tag{2.1}$$

Let $\hat{X}(t) = -X(t)$, then $(\hat{X}(t), B(t))$ is again a Markov additive process. Note that the level $\hat{X}(t)$ increases by i, $(i = 0, \pm 1)$ when $B(t)$ changes according to $\hat{Q}_i \equiv Q_{-i}$. This Markov additive process is referred to as a direction reversed process. For this additive process, we define matrices $\hat{N}_{k\ell}^{(0,+)}$ $\hat{N}_{k\ell}^{(-,m)}$ and $\hat{N}_{k\ell}^{(0,m)}$, $k, \ell \in \mathcal{N}$ and $\hat{R}$ similar to those for $(X(t), B(t))$. Obviously, we can see that

$$N_{k\ell}^{(0,+)} = \hat{N}_{(m-k)(m-\ell)}^{(-,m)}, \qquad N_{k\ell}^{(0,m)} = \hat{N}_{(m-k)(m-\ell)}^{(0,m)}.$$

Note that $\hat{R}$ is obtained as the minimal nonnegative solution for the matrix equation:

$$\hat{R}^2 Q_1 + \hat{R}Q_0 + Q_{-1} = O. \tag{2.2}$$

Let $\boldsymbol{\nu}$ be a stationary probability vector for $Q_{-1}+Q_0+Q_1$. Note that $\boldsymbol{\nu}$ is uniquely exists since $Q_{-1}+Q_0+Q_1$ is finite, irreducible and aperiodic. In this paper, we assume that

$$\beta \equiv \boldsymbol{\nu} Q_1 \mathbf{1} - \boldsymbol{\nu} Q_{-1} \mathbf{1} < 0, \tag{2.3}$$

unless stated otherwise. This condition represents that the mean drift of $X(t)$ under the stationary distribution $\boldsymbol{\nu}$ is negative, which implies that $X(t)$ goes to $-\infty$ w.p.1 as $t \to \infty$. We later put a reflecting boundary at the origin for the level process to be nonnegative, then (2.3) is necessary and sufficient for the reflected process to have the stationary distribution. In this case, (2.3) becomes a stability condition.

We now give some useful results for asymptotics of the occupation measures. As is well known, we have

$$N_{1m}^{(0,+)} = N_{1(m-1)}^{(0,+)} Q_1 N_{11}^{(0,+)}$$

by conditioning the last time when $X(t)$ leaves level $m-1$, so

$$N_{1m}^{(0,+)} = N_{11}^{(0,+)} R^{m-1}. \tag{2.4}$$

For example, see Theorem 6.2.7 of [11], in which (2.4) is obtained for the discrete time case. In what follows, we use the following assumption.

(i) The transition kernel $\{Q_\ell; \ell = 0, \pm 1\}$ is 1-arithmetic. That is, the greatest common divisor of the increments of the level process when the background process returns to the starting state is one. Precisely, $n \neq 0$ is such an increment for starting state $i \in \mathcal{S}_B$ if $P_i(X(t) - X(0) = n, B(t) = i) > 0$ for some $t > 0$. See [14] for a discrete time version of this definition.

This condition is naturally satisfied in queueing applications. From (2.4), asymptotic behaviors of $N_{1m}^{(0,+)}$ can be seen from those of R^m. The following proposition is a special case of Theorem 4.1 of [14], which answers to the latter behaviors.

Proposition 2.1 Assume that Markov additive process $(X(t), B(t))$ generated by Q_{-1}, Q_0 and Q_1 is 1-arithmetic. If $\beta < 0$, then there exist positive vectors $\boldsymbol{p}$, $\boldsymbol{q}$ and $\eta \in (0,1)$ such that

$$\boldsymbol{p}\left(Q_1 + \eta Q_0 + \eta^2 Q_{-1}\right) = \mathbf{0}, \tag{2.5}$$
$$\left(Q_1 + \eta Q_0 + \eta^2 Q_{-1}\right)\boldsymbol{q} = \mathbf{0}. \tag{2.6}$$

Then, the η is a Perron-Frobenius eigenvalue of R, and $\boldsymbol{p}$, $\boldsymbol{r}$ are the corresponding eigenvectors, where the vector $\boldsymbol{r}$ is given by

$$\boldsymbol{r} = -(Q_0 + (\eta I + R)Q_{-1})\boldsymbol{q}. \tag{2.7}$$

Hence, if assumption (i) holds, we have

$$\lim_{m\to\infty} \eta^{-m} R^m = \frac{1}{\boldsymbol{p}\boldsymbol{r}} \boldsymbol{r}\boldsymbol{p}. \tag{2.8}$$

Remark 2.1 Note that (2.1) is equivalent to

$$zQ_{-1} + Q_0 + z^{-1}Q_1 = (I - z^{-1}R)(Q_0 + (zI + R)Q_{-1}), \qquad z \neq 0. \tag{2.9}$$

From this and (2.6), we can see that $\boldsymbol{r}$ of (2.7) is the right eigenvector of R for eigenvalue η. Since η is the Perron-Frobenius eigenvalue of R, $\boldsymbol{r}$ must be nonnegative. (2.9) can be considered as a continuous time version of the Wiener-Hopf factorization for a discrete time Markov additive process with integer-valued level and finitely many background states (see, e.g., [1, 14]). □

We next put a reflection boundary at the origin for $(X(t), B(t))$ in such a way that $X(t) \geq 0$ and the transitions of $B(t)$ are changed only when $X(t)$ jumps into 0, stays at 0 or jumps out of 0. We denote this reflected process by $(Y(t), J(t))$, where level 0 has a background space $\mathcal{S}_0$, which may be different from $\mathcal{S}_B$. It is assumed that process $(Y(t), J(t))$ is a continuous time Markov chain with state space $\mathcal{S} \equiv \{0\} \times \mathcal{S}_0 \cup \mathcal{N}_+ \times \mathcal{S}_B$, where $\mathcal{N}_+ \equiv \{1, 2, \ldots\}$, and its transition rate matrix is given by the block matrix:

$$Q \equiv \begin{pmatrix} \underline{Q}_0 & \underline{Q}_1 & & & \\ \underline{Q}_{-1} & Q_0 & Q_1 & & \\ & Q_{-1} & Q_0 & Q_1 & \\ & & \ddots & \ddots & \ddots \\ & & & & \end{pmatrix},$$

where $\underline{Q}_0$, $\underline{Q}_1$ and $\underline{Q}_{-1}$ are $\mathcal{S}_0 \times \mathcal{S}_0$, $\mathcal{S}_0 \times \mathcal{S}_B$ and $\mathcal{S}_B \times \mathcal{S}_0$ matrices, respectively. In the literature, $(Y(t), B(t))$ is referred to as a quasi-birth-and-death process, QBD process for short. In the QBD process, the state of $Y(t)$ is called a level. Throughout the paper, we assume that Q is irreducible. To exclude a trivial case, we also assume that $\underline{Q}_1 \mathbf{1} \neq \mathbf{0}$.

Note that, for $k, \ell \geq 1$,

$$E\left[\int_0^{\tau^c_{(0,+)}} \boldsymbol{I}(Y(t) = \ell, J(t) = j)dt \middle| Y(0) = k, J(0) = i\right] = N^{(0,+)}_{k,\ell}(i, j), \quad i, j \in \mathcal{S}_B,$$

since $(X(t), B(t))$ is identical with $(Y(t), J(t))$ under the condition that $X(t), Y(t) \geq 1$. We here read $\tau^c_{(0,+)} = \inf\{t > 0; Y(t) \leq 0\}$ instead of $\tau^c_{(0,+)} = \inf\{t > 0; X(t) \leq 0\}$. This is abuse of the notation, but it simplifies notation. So, we use this kind of notation as long as there can be no confusion. It is well known that the stationary distribution of the QBD is obtained using R and therefore the occupation measures. This result is called a matrix geometric solution, and given by the following proposition (e.g., see Theorem 6.4.1 of [11]).

Proposition 2.2 If $\beta < 0$, then the QBD process $(Y(t), J(t))$ has a stationary distribution $\boldsymbol{\pi} \equiv (\boldsymbol{\pi}_\ell; \ell \geq 0)$, which is determined by

$$\boldsymbol{\pi}_0(\underline{Q}_0 + \underline{Q}_1 N^{(0,+)}_{11} \underline{Q}_{-1}) = \mathbf{0}, \tag{2.10}$$

$$\boldsymbol{\pi}_1 = \boldsymbol{\pi}_0 \underline{Q}_1 N^{(0,+)}_{11}, \tag{2.11}$$

$$\boldsymbol{\pi}_\ell = \boldsymbol{\pi}_1 R^{\ell-1}, \quad \ell \geq 2. \tag{2.12}$$

To compare the loss probability with the tail of stationary distribution, we note the following result, which is immediate from (2.8) and (2.12) (see also Theorem 4.1 of [14]).

Corollary 2.1 Under assumption (i) in addition to the assumptions of Proposition 2.2, we have

$$\lim_{n\to\infty} \eta^{-n}\boldsymbol{\pi}_n = \frac{\boldsymbol{\pi}_1\boldsymbol{r}}{\eta\boldsymbol{p}\boldsymbol{r}}\boldsymbol{p}. \tag{2.13}$$

3. A truncated QBD process

We next truncate the QBD process $(Y(t), J(t))$ so that $Y(t) \le m$, where m is a positive integer. We modify the transition rates of phase process $J(t)$ when $Y(t)$ changes from m to $m-1$ or stays at m. Then, we have a new continuous time Markov chain $(Y^{(m)}(t), J^{(m)}(t))$ with finite state space $\mathcal{N}_m \times \mathcal{S}_B$, where $\mathcal{N}_m \equiv \{0, 1, \ldots, m\}$. It is assumed that the process $(Y^{(m)}(t), J^{(m)}(t))$ has following level partitioned transition rate matrix:

$$Q^{(m)} \equiv \begin{pmatrix} \underline{Q}_0 & \underline{Q}_1 & & & & & \\ \underline{Q}_{-1} & Q_0 & Q_1 & & & & \\ & Q_{-1} & Q_0 & Q_1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & Q_{-1} & Q_0 & Q_1 & \\ & & & & Q_{-1} & Q_0 & Q_1 \\ & & & & & \overline{Q}_{-1} & \overline{Q}_0 \end{pmatrix},$$

where $\overline{Q}_{-1}$ and $\overline{Q}_0$ are $\mathcal{S}_B \times \mathcal{S}_B$ matrices. The process $(Y^{(m)}(t), J^{(m)}(t))$ is referred to as a truncated quasi-birth-and-death process, a truncated QBD process for short.

Remark 3.1 The set of the background states at level m is identical with those of the other level except for level 0, and the transition matrix from level $m-1$ to m are not modified. These are different from the modification at level 0. These modifications do not cause any problem in our queueing applications in Section 4, while it much simplifies our arguments. □

Using the last row block of $Q^{(m)}$, we can define a direction reversed QBD process $(\hat{Y}(t), \hat{J}(t))$ from the Markov additive process $(\hat{X}(t), B(t))$ in a similar way to define $(Y(t), J(t))$. Namely, this Markov chain $(\hat{Y}(t), \hat{J}(t))$ is assumed to have the following transition rate matrix:

$$\hat{Q} \equiv \begin{pmatrix} \overline{Q}_0 & \overline{Q}_{-1} & & & \\ Q_1 & Q_0 & Q_{-1} & & \\ & Q_1 & Q_0 & Q_{-1} & \\ & & \ddots & \ddots & \ddots \\ & & & & \end{pmatrix}.$$

Similarly to Q, we assume that transition rate matrix $\hat{Q}$ is irreducible throughout the paper.

Note that $Q^{(m)}$ has $(m+1)$ blocks in each row and column. We assume that $Q^{(m)}$ is irreducible and aperiodic, then $Q^{(m)}$ has a stationary probability vector $\boldsymbol{\pi}^{(m)} \equiv (\boldsymbol{\pi}_\ell^{(m)}; 0 \le \ell \le m)$, where $\boldsymbol{\pi}_\ell^{(m)}$ is subvector at level ℓ. Similarly to the case of the QBD, we have, for $1 \le k, \ell \le m-1$ and $i, j \in \mathcal{S}_B$,

$$\begin{aligned} E\left[\int_0^{\tau^c_{(0,m)}} \boldsymbol{I}(Y^{(m)}(t)=\ell, J^{(m)}(t)=j)dt \,\middle|\, Y^{(m)}(0)=k, J^{(m)}(0)=i\right] \\ = E_{(k,i)}\left[\int_0^{\tau^c_{(0,m)}} \boldsymbol{I}(X(t)=\ell, B(t)=j)dt\right] \\ = N_{k\ell}^{(0,m)}(i,j), \end{aligned}$$

since $(Y^{(m)}(t), J^{(m)}(t))$ is stochastically identical with $(X(t), B(t))$ when they take values in the set $(0,m) \times \mathcal{S}_B$. The following result is due to Hajek [8], which shows that the stationary distribution $\boldsymbol{\pi}^{(m)}$ is obtained through $N_{k\ell}^{(0,m)}$. For convenience of the reader, its proof is given in Appendix A.

Proposition 3.1 The stationary distribution $\boldsymbol{\pi}^{(m)} = (\boldsymbol{\pi}_\ell^{(m)}; 0 \le \ell \le m)$ for the truncated QBD process with transition rate matrix $Q^{(m)}$ is given by

$$\boldsymbol{\pi}_\ell^{(m)} = \boldsymbol{\pi}_0^{(m)} \underline{Q}_1 N_{1\ell}^{(0,m)} + \boldsymbol{\pi}_m^{(m)} \overline{Q}_{-1} N_{(m-1)\ell}^{(0,m)}, \quad (1 \le \ell \le m-1), \tag{3.1}$$

where $\boldsymbol{\pi}_0^{(m)}$ and $\boldsymbol{\pi}_{m-1}^{(m)}$ are determined by

$$\boldsymbol{\pi}_0^{(m)} \left(\underline{Q}_0 + \underline{Q}_1 N_{11}^{(0,m)} \underline{Q}_{-1}\right) + \boldsymbol{\pi}_m^{(m)} \overline{Q}_{-1} N_{(m-1)1}^{(0,m)} \underline{Q}_{-1} = \mathbf{0}, \tag{3.2}$$

$$\boldsymbol{\pi}_0^{(m)} \underline{Q}_1 N_{1(m-1)}^{(0,m)} Q_1 + \boldsymbol{\pi}_m^{(m)} \left(\overline{Q}_0 + \overline{Q}_{-1} N_{(m-1)(m-1)}^{(0,m)} Q_1\right) = \mathbf{0}, \tag{3.3}$$

It should be noticed that this proposition does not need any drift condition. Our major concern in this section is with an asymptotic behavior of $\boldsymbol{\pi}_m^{(m)}$ as $m \to \infty$ under the drift condition (2.3). This is answered by the following theorem.

Theorem 3.1 If $\beta < 0$, i.e., the negative drift condition (2.3) holds and if assumption (i) is satisfied, then we have

$$\lim_{m\to\infty} \eta^{-m} \boldsymbol{\pi}_m^{(m)} = \frac{\boldsymbol{\pi}_1 \boldsymbol{r}}{\eta \boldsymbol{p} \boldsymbol{r}} \boldsymbol{p} (\eta^{-1} I - \hat{R}) Q_1 (-\overline{U}_{00})^{-1}, \tag{3.4}$$

where η, $\boldsymbol{p}$ and $\boldsymbol{r}$ are determined by Proposition 2.1, and $\boldsymbol{\pi}_1$ and $\overline{U}_{00}$ are given by (2.11) and

$$\overline{U}_{00} = \overline{Q}_0 + \overline{Q}_{-1}(-Q_0 - \hat{R}Q_1)^{-1} Q_1.$$

Furthermore, the right-hand side of (3.4) is positive.

Remark 3.2 The decay rate η of $\boldsymbol{\pi}_m^{(m)}$ agrees with the one of $\boldsymbol{\pi}_m$ of the corresponding queue with an unlimited buffer. This is intuitively expected, and we are more interested in the prefactor of the geometric decay. □

Remark 3.3 It may look that R does not appear in (3.4). However, we generally need R to compute $\boldsymbol{\pi}_1\boldsymbol{r}$, while $\boldsymbol{p}\boldsymbol{r}$ can be computed from (2.7) as

$$\boldsymbol{p}\boldsymbol{r} = -\boldsymbol{p}(Q_0 + 2\eta Q_{-1})\boldsymbol{q},$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ can be computed directly from (2.5) and (2.6). □

Remark 3.4 Let $G = N_{11}^{(0,+)}Q_{-1}$ and $\hat{G} = \hat{N}_{11}^{(0,+)}Q_1$. These represents the hitting probabilities at the one lower levels of Markov additive process $(X(t), B(t))$ and its direction reversed process $(\hat{X}(t), B(t))$. So, $\hat{G}$ is a defective transition probability matrix by the drift condition (2.3). The following identities are immediate from the definitions.

$$RQ_{-1} = Q_1 G, \qquad \hat{R}Q_1 = Q_{-1}\hat{G}. \tag{3.5}$$

From the latter equation, it is easy to see that $\hat{R}$ can be used instead of $\hat{G}$ in computation of (3.4). We shall see that such a replacement is useful. As is well known, we have

$$\hat{N}_{11}^{(0,+)} = (-Q_0 - \hat{R}Q_1)^{-1}. \tag{3.6}$$

For example, see Proposition 6.4.2 of [11]. From this, we have

$$\overline{U}_{00} = \overline{Q}_0 + \overline{Q}_{-1}\hat{N}_{11}^{(0,+)}Q_1 = \overline{Q}_0 + \overline{Q}_{-1}\hat{G}.$$

Hence, $\overline{U}_{00}$ is a defective transition rate matrix, and it is indeed invertible. This is also intuitively clear since $\overline{U}_{00}$ is the transition rate for a censored process at level 0 of the direction reversed QBD $(\hat{Y}(t), \hat{J}(t))$. The latter is unstable, so the censored process is defective. □

Remark 3.5 Since (3.4) is new, the reader might wonder whether it is compatible with the existence results. Obviously, it is compatible with the corresponding result for the $M/M/1/m$ queue. We shall consider further compatibility in Remark 3.7. □

To prove this theorem, we give two lemmas, which may have independent interests.

Lemma 3.1 Under the assumptions of Theorem 3.1, we have

$$\lim_{m\to\infty} N_{k\ell}^{(0,m)} = N_{k\ell}^{(0,+)}, \tag{3.7}$$

$$\lim_{m\to\infty} N_{(m-1)(m-1)}^{(0,m)} = (-Q_0 - \hat{R}Q_1)^{-1}, \tag{3.8}$$

$$\lim_{m\to\infty} \eta^{-m} N_{1(m-1)}^{(0,m)} Q_1 = \frac{1}{\eta\boldsymbol{p}\boldsymbol{r}} N_{11}^{(0,+)}\boldsymbol{r}\boldsymbol{p}(-U_{00}), \tag{3.9}$$

where

$$U_{00} = Q_0 + Q_1 N_{11}^{(0,+)}Q_{-1} + Q_{-1}\hat{N}_{11}^{(0,+)}Q_1.$$

Furthermore, the right-hand side of (3.9) is a non-zero matrix, $\overline{U}_{00} = \overline{Q}_0 + \overline{Q}_{-1}\hat{N}_{11}^{(0,+)}Q_1$ is invertible, and $-\overline{U}_{00}^{-1}$ is positive.

PROOF. (3.7) is a direct consequence of the monotone convergence theorem. We next prove (3.8). Since $N^{(0,m)}_{(m-1)(m-1)} = \hat{N}^{(0,m)}_{11}$, we have

$$\lim_{m\to\infty} N^{(0,m)}_{(m-1)(m-1)} = \hat{N}^{(0,+)}_{11}$$

by the monotone convergence theorem. Hence, from (3.6), we get (3.8).

We next prove (3.9). By conditioning on the first visit of $(X(t), B(t))$ to level m before hitting level 0, we have

$$N^{(0,+)}_{1m} = N^{(0,m)}_{1(m-1)} Q_1 N^{(0,+)}_{mm}. \tag{3.10}$$

Since this equation is a key in our arguments, we also give its formal proof in Appendix B. Combining this with (2.4), we have

$$\begin{aligned} N^{(0,+)}_{11} R^{m-1} &= N^{(0,m)}_{1(m-1)} Q_1 N^{(0,+)}_{mm} \\ &= N^{(0,m)}_{1(m-1)} Q_1 \hat{N}^{(-,m)}_{00}. \end{aligned}$$

Hence, Proposition 2.1 and the monotone convergence of $\hat{N}^{(-,m)}_{00}$ to $\hat{N}_{00}$ yield

$$\lim_{m\to\infty} \eta^{-m} N^{(0,m)}_{1(m-1)} Q_1 \hat{N}_{00} = \frac{1}{\eta \boldsymbol{p}\boldsymbol{r}} N^{(0,+)}_{11} \boldsymbol{r}\boldsymbol{p}, \tag{3.11}$$

where

$$\hat{N}_{00}(i,j) = E_{(0,i)}\left[\int_0^\infty \boldsymbol{I}(\hat{X}(t) = 0, B(t) = j)dt\right], \qquad i,j \in \mathcal{S}_B.$$

Note that

$$U_{00} = Q_0 + Q_1 G + Q_{-1}\hat{G},$$

and U_{00} is the transition rate matrix of $(\hat{X}(t), B(t))$ censoring at level 0. Furthermore, it is invertible since $\hat{G}$ is defective as mentioned in Remark 3.4. So, we have

$$\hat{N}_{00} = (-U_{00})^{-1}.$$

This and (3.11) conclude (3.9). Assume that the right-hand side of (3.9) is the null matrix. Since U_{00} is invertible, this implies that

$$N^{(0,+)}_{11} \boldsymbol{r}\boldsymbol{p} = 0.$$

Similar to $\hat{N}^{(0,+)}_{11}$ in (3.6), $N^{(0,+)}_{11}$ is invertible, so we have $\boldsymbol{r}\boldsymbol{p} = 0$. This contradicts the fact that $\boldsymbol{p}$ is positive and $\boldsymbol{r}$ is not null. Hence, the right-hand side of (3.9) can not be null.

Since $\hat{N}^{(0,+)}_{11} Q_1$ is defective, $\overline{U}_{00}$ is a defective transition rate matrix, so it is invertible. Since $\hat{Q}$ is irreducible, $\overline{U}_{00}$ must be irreducible. Hence, $-\overline{U}^{-1}_{00}$ is positive. □

Lemma 3.2 Under the assumptions of Theorem 3.1, we have

$$\lim_{m\to\infty} \boldsymbol{\pi}_\ell^{(m)} = \boldsymbol{\pi}_\ell, \qquad \ell \geq 0.$$

PROOF. Note that $Q^{(m)}(\boldsymbol{i},\boldsymbol{j})$ converges to $Q(\boldsymbol{i},\boldsymbol{j})$ as $m \to \infty$ for each $\boldsymbol{i},\boldsymbol{j} \in \mathcal{S}$ and Q has the stationary distribution. Hence, by Lemma 2.1 of [21], we only need to show that $\boldsymbol{\pi}^{(m)} \equiv (\boldsymbol{\pi}_\ell^{(m)}; \ell \in \mathcal{N}_m)$ is tight for $m \geq 1$, that is, for each $\epsilon > 0$, there exists a positive integer m_0 such that

$$\sum_{\ell=m_0}^{m} \boldsymbol{\pi}_\ell^{(m)} \mathbf{1} < \epsilon, \qquad \forall m \geq m_0,$$

Since

$$N_{(m-1)(m-1)}^{(0,m)} Q_1 \leq N_{(m-1)(m-1)}^{(-,m)} Q_1 = \hat{N}_{11}^{(0,+)} Q_1$$

and $\hat{N}_{11}^{(0,+)} Q_1$ is defective, (3.3) yields

$$\boldsymbol{\pi}_m^{(m)} \leq \boldsymbol{\pi}_0^{(m)} \underline{Q}_1 N_{11}^{(0,+)} R^{m-2} Q_1 \left\{ -\left(\overline{Q}_0 + \overline{Q}_{-1} \hat{N}_{11}^{(0,+)} Q_1 \right)^{-1} \right\}.$$

Using this and the fact that

$$N_{(m-1)\ell}^{(0,m)} \leq \hat{N}_{1(m-\ell)}^{(0,+)} = \hat{N}_{11}^{(0,+)} \hat{R}^{m-\ell-1},$$

we have, from (3.1), for $1 \leq \ell \leq m-1$,

$$\boldsymbol{\pi}_\ell^{(m)} \leq \boldsymbol{\pi}_0^{(m)} \underline{Q}_1 N_{11}^{(0,+)} \left(R^{\ell-1} - R^{m-2} Q_1 \left(\overline{Q}_0 + \overline{Q}_{-1} \hat{N}_{11}^{(0,+)} Q_1 \right)^{-1} \overline{Q}_{-1} \hat{N}_{11}^{(0,+)} \hat{R}^{m-\ell-1} \right),$$

Since the Perron-Frobenius eigenvalue of $\hat{R}$ is unit, $\hat{R}^\ell$ is uniformly bounded in ℓ. On the other hand, R^ℓ geometrically decay with rate $\eta < 1$ by (2.8). Since all the entries of $\boldsymbol{\pi}_0^{(m)}$ are bounded by 1, the above two inequalities conclude the required tightness. □

Remark 3.6 Lemma 3.2 can be also obtained from Theorem 3.4 of [12] using uniformization of transition rate matrices Q and $Q^{(m)}$, since the uniformized probability matrices P and $P^{(m)}$ corresponding to Q and $Q^{(m)}$ are stochastically block-monotone and $P^{(m)}$ is a north-west corner truncation and block-augmented matrix of P.

PROOF OF THEOREM 3.1 From (3.3), we have

$$\boldsymbol{\pi}_m^{(m)} = \boldsymbol{\pi}_0^{(m)} \underline{Q}_1 N_{1(m-1)}^{(0,m)} Q_1 \left(-\overline{Q}_0 - \overline{Q}_{-1} N_{(m-1)(m-1)}^{(0,m)} Q_1 \right)^{-1}.$$

Multiplying both sides of this equation with η^{-m} and letting $m \to \infty$, Lemmas 3.1 and Lemma 3.2 and (2.11) yield

$$\lim_{m\to\infty} \eta^{-m} \boldsymbol{\pi}_m^{(m)} = \frac{\boldsymbol{\pi}_1 \boldsymbol{r}}{\eta \boldsymbol{p} \boldsymbol{r}} \boldsymbol{p} U_{00} \overline{U}_{00}^{-1},$$

where the positivity of the right-hand side of this equation is immediate from the last statement of Lemma 3.1. Thus, (3.4) is obtained if we show that

$$\boldsymbol{p}U_{00} = -\boldsymbol{p}(\eta^{-1}I - \hat{R})Q_1.$$

Since $\boldsymbol{p}$ is the left invariant vector of R and satisfies (2.5), we have

$$\begin{aligned} \boldsymbol{p}U_{00} &= \boldsymbol{p}(Q_0 + RQ_{-1} + \hat{R}Q_1) \\ &= \boldsymbol{p}(Q_0 + \eta Q_{-1} + \hat{R}Q_1) \\ &= \boldsymbol{p}(-\eta^{-1}Q_1 + \hat{R}Q_1). \end{aligned}$$

Thus, we get the desired result. □

The following result shows that the constant vector in (3.4) can be simplified under some assumptions on transitions at the upper boundary. These assumptions are typical in queueing applications, and will be used in the next section.

Corollary 3.1 Under the assumptions of Theorem 3.1, if $\overline{Q}_0 = Q_0 + Q_1$ and $\overline{Q}_{-1} = Q_{-1}$, then we have

$$\lim_{m\to\infty} \eta^{-m}\boldsymbol{\pi}_m^{(m)} = \frac{\boldsymbol{\pi}_1\boldsymbol{r}}{\eta(1-\eta)\boldsymbol{p}\boldsymbol{r}}\boldsymbol{p}(I - \eta\hat{R}). \tag{3.12}$$

Proof. Similarly to (2.9), (2.2) is equivalent to

$$zQ_1 + Q_0 + z^{-1}Q_{-1} = (I - z^{-1}\hat{R})(Q_0 + \hat{R}Q_1 + zQ_1), \qquad z \neq 0. \tag{3.13}$$

Letting $z = \eta^{-1}$ in this equation and multiplying $\boldsymbol{p}$ from the left, we have, from (2.5),

$$\boldsymbol{p}(\eta^{-1}I - \hat{R})(\eta(Q_0 + \hat{R}Q_1) + Q_1) = \mathbf{0}.$$

This implies that

$$\begin{aligned} (1-\eta)\boldsymbol{p}(\eta^{-1}I - \hat{R})Q_1 &= -\eta\boldsymbol{p}(\eta^{-1}I - \hat{R})(Q_0 + Q_1 + \hat{R}Q_1) \\ &= \eta\boldsymbol{p}(\eta^{-1}I - \hat{R})(-\overline{U}_{00}). \end{aligned}$$

Applying this to (3.4), we have (3.12). □

Remark 3.7 As far as the authors know, (3.4) and (3.12) have not been known even for special cases except trivial models such as the $M/M/1/m$ queue. However, we can produce a less trivial example for this using the $M/PH/1/m$ queue. In this model, the service times are subject to a phase type distribution. Denote this distribution by $PH(T, \boldsymbol{\alpha})$ (see Example 4.1 for its definition). Then, this queue is described by the QBD with $Q_1 = \lambda I$, $Q_0 = T - \lambda I$ and $Q_{-1} = \boldsymbol{t}\boldsymbol{\alpha}$, where λ is the mean arrival rate and $\boldsymbol{t} = (-T)\mathbf{1}$. Its stationary joint distribution is known (see Section 3.2 of [15]). From this result, we have

$$\boldsymbol{\pi}_m^{(m)} = \lambda\boldsymbol{\pi}_1 R^{m-2}(-T)^{-1},$$

where $R = \lambda(-T + \lambda I - \lambda\boldsymbol{t}\boldsymbol{\alpha})^{-1}$. This implies

$$\lim_{m\to\infty} \eta^{-m}\boldsymbol{\pi}_m^{(m)} = \frac{\lambda\boldsymbol{\pi}_1\boldsymbol{r}}{\eta^2\boldsymbol{p}\boldsymbol{r}}\boldsymbol{p}(-T)^{-1}, \tag{3.14}$$

In Appendix C, we show that (3.14) indeed agrees with (3.12) for the $M/PH/1/m$ queue. □

One may wonder whether vector $\boldsymbol{\pi}_m^{(m)}$ can be proportional to $\boldsymbol{\pi}_m$ as m goes to infinity under the setting of Corollary 3.1. This is answered by the following corollary.

Corollary 3.2 Under the assumptions of Corollary 3.1, the right-hand sides of (2.13) and (3.12) are proportional if and only if $\boldsymbol{p}$ is proportional to $\boldsymbol{\nu}$. In this case, the limits in (2.13) and (3.12) are identical, and we always have $\eta = \rho$, where $\rho = \boldsymbol{\nu}Q_1\mathbf{1}/(\boldsymbol{\nu}Q_{-1}\mathbf{1})$.

PROOF. By Corollary 3.1, the proportionality holds if and only if we have, for some $c > 0$,

$$\boldsymbol{p}(I - \eta\hat{R}) = c(1-\eta)\boldsymbol{p}. \tag{3.15}$$

Note that the Perron-Frobenius eigenvalue of $\hat{R}$ is 1. Denote the associated right eigenvector by $\hat{\boldsymbol{r}}$, and postmultiply both sides of (3.15) with it. Then, c must be 1. So, (3.15) implies $\boldsymbol{p}\hat{R} = \boldsymbol{p}$. However, $\boldsymbol{\nu}$ is the left invariant vector of $\hat{R}$, which easily follows from (3.13) with $z = 1$. Hence, $\boldsymbol{p}$ must be proportional to $\boldsymbol{\nu}$. In turn, if this proportionality holds, then we have (3.15) with $c = 1$, and the two limits are identical. Furthermore, we have

$$\boldsymbol{\nu}(Q_1 + \eta Q_0 + \eta^2 Q_{-1}) = \mathbf{0}, \tag{3.16}$$

Since $\boldsymbol{\nu}$ is the stationary distribution, this yields

$$(1-\eta)\boldsymbol{\nu}(Q_1 - \eta Q_{-1}) = \mathbf{0},$$

Hence, η must be ρ. □

Remark 3.8 Consider Q_{-1}, Q_0 and Q_1 such that $Q_1 = aQ_{-1}$ for some positive $a < 1$. Then, ρ must be a, and we have

$$\boldsymbol{\nu}(aQ_{-1} + Q_0 + Q_{-1}) = \mathbf{0}.$$

Multiplying this with a and using $Q_1 = aQ_{-1}$ again yield (3.16) with $\eta = a$. Hence, we have $\boldsymbol{p} = \boldsymbol{\nu}$. Thus, we have the example that satisfies $\boldsymbol{p} = \boldsymbol{\nu}$. However, this is not a queueing type model, which is discussed in the next section. Furthermore, $\eta = \rho$ is hardly expected for queueing applications except the case of a single background state. So, the situation of Corollary 3.2 is hard to happen in them.

In Theorem 3.1, we have assumes that $\beta < 0$. However, the other two cases $\beta = 0$ and $\beta > 0$ may be interesting. For $\beta > 1$, the following result is immediate from the reversed direction version of Lemma 3.2.

Corollary 3.3 If $\beta > 0$, then the direction reversed QBD with transition rate matrix $\hat{Q}$ has the stationary distribution $\hat{\boldsymbol{\pi}} \equiv (\hat{\boldsymbol{\pi}}_\ell; \ell \geq 0)$, and

$$\lim_{m\to\infty} \boldsymbol{\pi}_m^{(m)} = \hat{\boldsymbol{\pi}}_0. \tag{3.17}$$

If $\beta = 0$, then

$$\lim_{m\to\infty} \boldsymbol{\pi}_m^{(m)} = \mathbf{0}. \tag{3.18}$$

In [4], more detailed limiting behaviors are obtained for the loss probability of $MArP/G/1/K+1$ queue. The corresponding results may be expected for the boundary probabilities. For example, we may conjecture that, if $\beta = 0$, then

$$\lim_{m\to\infty} m\boldsymbol{\pi}_m^{(m)} = \boldsymbol{c}$$

for some positive constant vector $\boldsymbol{c}$.

4. A $MArP/MSP/c$ with a finite buffer

We consider a $MArP/MSP$ queue with a finite buffer of sizes $K \geq 0$. This queue and its variants have been studied under the name $MAP/MSP/1$ in [16]. Here, MArP represents Markovian arrival process which is termed by Neuts. Similarly, MSP represents Markovian service process, which is another MArP whose arrivals are replaced with service completions. This MArP (MSP) is a counting process controlled by a underlying continuous time Markov chain. Note that the number of servers as well as service discipline are not essential as long as service does not depend on the number of customers in system. The difference only appears at the lower boundary. Just for convenience, we assume that there are c servers, and denote the model by $MArP/MSP/c/K+c$. In this model, level 0 accommodate all the states in which the number of customers in system are not greater than $c-1$, so level $\ell \geq 1$ represents the states that there are $\ell+c-1$ customers in system. It is also assumed that arrival and service processes are independent of each other as long as all servers are busy. In this way, we describe the $MArP/MSP/c/K+c$ queue by the QBD process truncated by level $K+1$.

Let $m = K+1$ and let $L^{(m)}(t)$ be the number of customers in system at time t. We define the level at time t by

$$Y^{(m)}(t) = \max(0, L^{(m)}(t) - c + 1).$$

Then, we can formulate $MArP/MSP/c/K+c$ as a truncated QBD with level $Y^{(m)}(t)$. To describe this QBD, denote background processes for arrivals and services by $\{J_1^{(m)}(t)\}_{t\geq 0}$ and $\{J_2^{(m)}(t)\}_{t\geq 0}$, respectively. Process $\{J_1^{(m)}(t)\}$ is a continuous time Markov chain with finite state space $\mathcal{M}_1 \equiv \{1, 2, \ldots, k_1\}$. However, $J_2^{(m)}(t)$ may depend on $L^{(m)}(t)$ and $J_1^{(m)}(t)$, and it is activated only when there are customers in system. Thus, $\{J_2^{(m)}(t)\}$ is not a Markov chain, but it becomes so when $Y^{(m)}(t)$ remains in the set $\{1, 2, \ldots, m\}$. Let $\mathcal{M}_2 \equiv \{1, 2, \ldots, k_2\}$ be a state space of this Markov chain. Denote the transition rate matrix of these Markov chains by $k_1 \times k_1$ ($k_2 \times k_2$) matrix $C_1 + D_1$ ($C_2 + D_2$), respectively, where C_i is a defective rate matrix, and D_i is a nonnegative matrix for $i = 1, 2$. If transitions due to D_1 (D_2) occur, then a customer arrives (a service is completed).

Thus, C_1 and D_1 (C_2 and D_2) generate a MArP (MSP when all servers are busy). It is assumed that $C_1 + D_1$ ($C_2 + D_2$) is irreducible and aperiodic. This implies that there is a unique stationary probability vector $\boldsymbol{\nu}_1$ ($\boldsymbol{\nu}_2$) satisfying $\boldsymbol{\nu}_1(C_1+D_1) = \mathbf{0}$ ($\boldsymbol{\nu}_2(C_2+D_2) = \mathbf{0}$).

When $Y^{(m)}(t)$ attains value 0 either after or before its state change, transitions of $J_2^{(m)}(t)$ depends on $L^{(m)}(t)$ and $J_1^{(m)}(t)$, where $L^{(m)}(t)$ takes values in $\{0, 1, \ldots, c\}$, and we need them to describe a service process. Thus, the background process can be given

by $(L^{(m)}(t), J_1^{(m)}(t), J_2^{(m)}(t))$ during $Y^{(m)}(t) = 0$. Denote its state space by $\mathcal{S}_0$. We then consider three cases that $Y_2^{(m)}$ changes from 1 to 0, 0 to 0 and 0 to 1, separately. For these cases, we denote the transition rate matrices of the background process by $\underline{Q}_{-1}$, $\underline{Q}_0$ and $\underline{Q}_1$, respectively. It is routine to give detailed descriptions for them (e.g, see [16] and Example 4.2 below for the case of $c = 1$). However, we omit those descriptions since our arguments does not depend on their specific forms.

Let

$$Z^{(m)}(t) \equiv \left(Y^{(m)}(t), J_1^{(m)}(t), \underline{J}_2^{(m)}(t)\right),$$

where $\underline{J}_2^{(m)}(t) = (L^{(m)}(t), J_2^{(m)}(t))$ when $Y^{(m)}(t) = 0$, and $\underline{J}_2^{(m)}(t) = J_2^{(m)}(t)$ otherwise. Then, $Z^{(m)}(t)$ is a Markov chain with state space $\{0\} \times \mathcal{S}_0 \cup \{1, 2, \ldots, m\} \times \mathcal{S}_B$, where $\mathcal{S}_B \equiv \mathcal{M}_1 \times \mathcal{M}_2$. Note that we can set $\underline{J}_2^{(m)}(t) = J_2^{(m)}(t)$ for $c = 1$. Referring to $Y^{(m)}(t)$ as a level process and to $(J_1^{(m)}(t), \underline{J}_2^{(m)}(t))$ as a background process, the Markov chain $Z^{(m)}(t)$ is a truncated QBD process with a transition rate matrix $Q^{(m)}$, whose blocks are given by

$$\begin{aligned} Q_0 &= C_1 \oplus C_2, \qquad \overline{Q}_0 = (C_1 + D_1) \oplus C_2, \\ Q_{-1} &= \overline{Q}_{-1} = I_1 \otimes D_2, \qquad Q_1 = D_1 \otimes I_2, \end{aligned}$$

where $\otimes$ and $\oplus$ stand for the Kronecker product and sum (e.g., see [2]), and I_1, I_2 are identity matrices of appropriate sizes. For matrices $\underline{Q}_{-1}, \underline{Q}_0$ and $\underline{Q}_1$, no specific forms are given as we discussed above.

It is easy to see that the stationary probability vector for $Q_{-1} + Q_0 + Q_1$ is $\boldsymbol{\nu}_1 \otimes \boldsymbol{\nu}_2$. Let

$$\lambda = \boldsymbol{\nu}_1 D_1 \mathbf{1}, \qquad \mu = \frac{1}{c} \boldsymbol{\nu}_2 D_2 \mathbf{1}.$$

Clearly, λ is the mean arrival rate of customers, and μ is the mean service rate at each server. Then, in this model, the negative drift condition (2.3) is equivalent to

$$\rho \equiv \frac{\lambda}{c\mu} < 1. \tag{4.1}$$

Note that this queue always has the stationary distribution since levels are finite.

Compute the matrix generating function $Q_1 + zQ_0 + z^2 Q_{-1}$ for real number z. Then, we have

$$Q_1 + zQ_0 + z^2 Q_{-1} = (zC_1 + D_1) \oplus (zC_2 + z^2 D_2). \tag{4.2}$$

By the Perron-Frobenius theorem, for each $z \geq 0$, there exists positive eigenvectors $\boldsymbol{q}_1(z)$, $\boldsymbol{q}_2(z)$, $\boldsymbol{p}_1(z)$, $\boldsymbol{p}_2(z)$ and the corresponding nonnegative eigenvalues $\theta_1(z)$, $\theta_2(z)$ such that

$$\begin{aligned} \boldsymbol{p}_1(z)(zC_1 + D_1) &= \theta_1(z)\boldsymbol{p}_1(z), \qquad \boldsymbol{p}_2(z)(zC_2 + z^2 D_2) = \theta_2(z)\boldsymbol{p}_2(z), & (4.3) \\ (zC_1 + D_1)\boldsymbol{q}_1(z) &= \theta_1(z)\boldsymbol{q}_1(z), \qquad (zC_2 + z^2 D_2)\boldsymbol{q}_2(z) = \theta_2(z)\boldsymbol{q}_2(z). & (4.4) \end{aligned}$$

Applying these formulas to (4.2), we have

$$\begin{aligned} \boldsymbol{p}_1(z) \otimes \boldsymbol{p}_2(z) \left(Q_1 + zQ_0 + z^2 Q_{-1}\right) &= (\theta_1(z) + \theta_2(z))\, \boldsymbol{p}_1(z) \otimes \boldsymbol{p}_2(z), & (4.5) \\ \left(Q_1 + zQ_0 + z^2 Q_{-1}\right) \boldsymbol{q}_1(z) \otimes \boldsymbol{q}_2(z) &= (\theta_1(z) + \theta_2(z))\, \boldsymbol{q}_1(z) \otimes \boldsymbol{q}_2(z). & (4.6) \end{aligned}$$

Lemma 4.1 Under the stability condition (4.1), there exists a unique $\eta \in (0,1)$ such that

$$\theta_1(\eta) + \theta_2(\eta) = 0. \tag{4.7}$$

For this η, let $\boldsymbol{p} = \boldsymbol{p}_1(\eta) \otimes \boldsymbol{p}_2(\eta)$ and $\boldsymbol{q} = \boldsymbol{q}_1(\eta) \otimes \boldsymbol{q}_2(\eta)$, then these $\boldsymbol{p}$ and $\boldsymbol{q}$ satisfy (2.5) and (2.6), that is,

$$\boldsymbol{p}_1(\eta) \otimes \boldsymbol{p}_2(\eta) \left(Q_1 + \eta Q_0 + \eta^2 Q_{-1}\right) = \boldsymbol{0}, \tag{4.8}$$
$$\left(Q_1 + \eta Q_0 + \eta^2 Q_{-1}\right) \boldsymbol{q}_1(\eta) \otimes \boldsymbol{q}_2(\eta) = \boldsymbol{0}. \tag{4.9}$$

PROOF. Since $\theta_1(z)$ and $\theta_2(z)$ are convex (see e.g., [10]), $\theta_1(z) + \theta_2(z)$ is also convex. On the other hand, $\theta_1'(1) = -\boldsymbol{\nu}_1 D_1 \mathbf{1}$, and $\theta_2'(1) = \boldsymbol{\nu}_2 D_2 \mathbf{1}$, so (4.1) implies that

$$\theta_1'(1) + \theta_2'(1) > 0.$$

Since $\theta_1(0) > 0$, there must be a unique $\eta \in (0,1)$ satisfying (4.7). Equations (4.8) and (4.9) are immediate from (4.5), (4.6) and (4.7). □

We are now ready to consider the loss probability $p_{\text{loss}}^{(K+c)}$ of the $MArP/MSP/c/K+c$ queue, which is defined as

$$p_{\text{loss}}^{(K+c)} = \lambda^{-1} \boldsymbol{\pi}_m^{(m)} Q_1 \mathbf{1},$$

where $m = K + 1$.

Theorem 4.1 For the $MArP/MSP/c/K+c$ queue, if $\rho < 1$, then we have

$$\lim_{K\to\infty} \eta^{-(K+c)} p_{\text{loss}}^{(K+c)} = \frac{(1-\rho)\boldsymbol{\pi}_c^{(\text{queue})} \boldsymbol{r} \boldsymbol{p} \hat{\boldsymbol{r}}}{\rho \eta^{c-1} (1-\eta) \boldsymbol{p} \boldsymbol{r}}, \tag{4.10}$$

where $\boldsymbol{\pi}_c^{(\text{queue})} = \boldsymbol{\pi}_1$, i.e., it is the positive vector for the stationary probabilities that there are c customers in system, $\eta \in (0,1)$ is the solution of (4.7) and

$$\boldsymbol{p} = \boldsymbol{p}_1(\eta) \otimes \boldsymbol{p}_2(\eta),$$
$$\boldsymbol{r} = -(Q_0 + (\eta I + R) Q_{-1}) \boldsymbol{q}_1(\eta) \otimes \boldsymbol{q}_2(\eta),$$

and $\hat{\boldsymbol{r}}$ is the right invariant vector of $\hat{R}$ normalized as $\boldsymbol{\nu} \hat{\boldsymbol{r}} = 1$, which is given by

$$\hat{\boldsymbol{r}} = -(c\mu - \lambda)^{-1} (Q_0 + Q_1 + \hat{R} Q_1) \mathbf{1}.$$

Remark 4.1 It is interesting to see how the loss probability is different from the stationary probability at level $K+1$, i.e., $p_{K+c}^{(\infty)} \equiv \boldsymbol{\pi}_{K+1} \mathbf{1}$, of the corresponding unlimited buffer queue. They have the same decay rate, and their ratio is given by

$$\lim_{K\to\infty} \frac{p_{\text{loss}}^{(K+c)}}{p_{K+c}^{(\infty)}} = \frac{\eta(1-\rho)\boldsymbol{p}\hat{\boldsymbol{r}}}{\rho(1-\eta)\boldsymbol{p}\mathbf{1}}.$$

This may be useful for estimating $p_{\text{loss}}^{(K+c)}$ for large K if we know the exact values of $p_{K+c}^{(\infty)}$. □

PROOF. From Corollary 3.1 and Lemma 4.1, (4.10) is obtained if we have

$$\boldsymbol{p}(I - \eta\hat{R})Q_1\mathbf{1} = \eta c\mu(1-\rho)\boldsymbol{p}\hat{\boldsymbol{r}}. \tag{4.11}$$

We first show that $\hat{\boldsymbol{r}}$ is the right invariant vector of $\hat{R}$ satisfying the normalization. From (3.13) with $z = 1$, we have

$$\mathbf{0} = (Q_1 + Q_0 + Q_{-1})\mathbf{1} = (I - \hat{R})(Q_0 + Q_1 + \hat{R}Q_1)\mathbf{1}. \tag{4.12}$$

Hence, $\hat{\boldsymbol{r}}$ is the right invariant of $\hat{R}$, where the minus sign is need for $\hat{\boldsymbol{r}}$ to be positive. Similarly, $\boldsymbol{\nu}$ is the left invariant vector. This implies that

$$\begin{aligned} \boldsymbol{\nu}\hat{\boldsymbol{r}} &= -(c\mu - \lambda)^{-1}\boldsymbol{\nu}(Q_0 + Q_1 + Q_1)\mathbf{1} \\ &= (c\mu - \lambda)^{-1}\boldsymbol{\nu}(Q_{-1} - Q_1)\mathbf{1} = 1. \end{aligned}$$

Also, from (4.12), we have

$$\hat{R}Q_1\mathbf{1} = -(Q_0 + Q_1)\mathbf{1} - (c\mu - \lambda)\hat{\boldsymbol{r}} = Q_{-1}\mathbf{1} - (c\mu - \lambda)\hat{\boldsymbol{r}}.$$

This yields

$$\boldsymbol{p}(I - \eta\hat{R})Q_1\mathbf{1} = \boldsymbol{p}(Q_1 - \eta Q_{-1})\mathbf{1} + \eta(c\mu - \lambda)\boldsymbol{p}\hat{\boldsymbol{r}}.$$

Hence, we get (4.11) if $\boldsymbol{p}(Q_1 - \eta Q_{-1})\mathbf{1} = 0$. To prove this, we first derive the following formulas from (4.3) using the fact that $(C_i + D_i)\mathbf{1} = \mathbf{0}$.

$$\begin{aligned} (1-\eta)\boldsymbol{p}_1(\eta)D_1\mathbf{1}_1 &= \theta_1(\eta)\boldsymbol{p}_1(\eta)\mathbf{1}_1, \\ \eta(1-\eta)\boldsymbol{p}_2(\eta)D_2\mathbf{1}_2 &= -\theta_2(\eta)\boldsymbol{p}_2(\eta)\mathbf{1}_2. \end{aligned}$$

These yield

$$\begin{aligned} \boldsymbol{p}(Q_1 - \eta Q_{-1})\mathbf{1} &= \boldsymbol{p}_1(\eta) \otimes \boldsymbol{p}_2(\eta)(D_1 \otimes I_2 - \eta I_1 \otimes D_2)\mathbf{1} \\ &= \frac{1}{1-\eta}(\theta_1(\eta) + \theta_2(\eta))\boldsymbol{p}_1(\eta)\mathbf{1}_1\boldsymbol{p}_2(\eta)\mathbf{1}_2 = 0. \end{aligned}$$

Thus, (4.11) is obtained, which concludes (4.10). □

In applications of Theorem 4.1, we need to compute $\eta, \boldsymbol{\pi}_c^{\text{queue}}, \boldsymbol{p}, \boldsymbol{r}$ and $\hat{\boldsymbol{r}}$. In principle, these can be done through computing R and $\hat{R}$. However, their analytic expressions may be also interesting from the theoretical point of view. The following result partially answers to them.

Corollary 4.1 Under the assumptions and normalization of Theorem 4.1, if $\boldsymbol{p}_i(\eta)$ and $\boldsymbol{q}_i(\eta)$ are normalized in such a way that

$$\boldsymbol{p}_i(\eta)\mathbf{1}_i = 1, \qquad \boldsymbol{p}_i(\eta)\boldsymbol{q}_i(\eta) = 1, \qquad i = 1, 2, \tag{4.13}$$

then

$$\lim_{K\to\infty} \eta^{-(K+c)} p_{\text{loss}}^{(K+c)} = \frac{(1-\rho)\boldsymbol{\pi}_c^{(\text{queue})}\boldsymbol{r}\boldsymbol{p}\hat{\boldsymbol{r}}}{\rho\eta^{c-1}(1-\eta)(-(\theta_1'(\eta) + \theta_2'(\eta)))}, \tag{4.14}$$

PROOF. Differentiating (4.3) and (4.4) with respect to z at $z = \eta$ and postmultiplying with $\boldsymbol{q}_1(\eta)$ and $\boldsymbol{q}_2(\eta)$, respectively, we have

$$\begin{aligned}
&\boldsymbol{p}_1(\eta)C_1\boldsymbol{q}_1(\eta) = \theta_1'(\eta)\boldsymbol{p}_1(\eta)\boldsymbol{q}_1(\eta) = \theta_1'(\eta),\\
&\boldsymbol{p}_2(\eta)(C_2 + 2\eta D_2)\boldsymbol{q}_2(\eta) = \theta_2'(\eta)\boldsymbol{p}_2(\eta)\boldsymbol{q}_2(\eta) = \theta_2'(\eta),
\end{aligned}$$

where the second equalities follow from the normalization. Using these formulas and the normalization (4.13), we have

$$\begin{aligned}
\boldsymbol{pr} &= -\boldsymbol{p}_1(\eta)\otimes\boldsymbol{p}_2(\eta)\left(C_1\otimes I_2 + I_1\otimes C_2 + 2\eta I_1\otimes D_2\right)\boldsymbol{q}_1(\eta)\otimes\boldsymbol{q}_2(\eta)\\
&= -\big(\boldsymbol{p}_1(\eta)C_1\boldsymbol{q}_1(\eta)\cdot\boldsymbol{p}_2(\eta)\boldsymbol{q}_2(\eta)\\
&\qquad +\boldsymbol{p}_1(\eta)\boldsymbol{q}_1(\eta)\cdot\boldsymbol{p}_2(\eta)C_2\boldsymbol{q}_2(\eta) + 2\eta\boldsymbol{p}_1(\eta)\boldsymbol{q}_1(\eta)\cdot\boldsymbol{p}_2(\eta)D_2\boldsymbol{q}_2(\eta)\big)\\
&= -\left(\theta_1'(\eta) + \theta_2'(\eta)\right). \qquad (4.15)
\end{aligned}$$

Substituting this into (4.10), we have (4.14). □

Example 4.1 (Phase type renewal process) A simple but important example for MArP and MSP is a renewal process with a phase type interarrival distribution. Suppose that transition rate matrix T and nonnegative matrix D generate a MArP. Let $D = \boldsymbol{t}\boldsymbol{\alpha}$ for a probability vector $\boldsymbol{\alpha}$, where $\boldsymbol{t} = (-T)\mathbf{1}$. Then, this MArP becomes a renewal process with an interarrival distribution whose density f is given by

$$f(x) = \boldsymbol{\alpha}\exp(xT)\boldsymbol{t}, \qquad x \geq 0.$$

This distribution is called a phase type, and denoted by $PH(\boldsymbol{\alpha}, T)$ (e.g., see [11]). We refer to this renewal process as a phase type renewal process. For the phase type renewal process generated by $PH(\boldsymbol{\alpha}, T)$, (4.3) and (4.4) become simpler. For example, let $\theta(z), \boldsymbol{p}(z)$ and $\boldsymbol{q}(z)$ are the Perron-Frobenius eigenvalue and left and right eigenvectors of $zT + \boldsymbol{t}\boldsymbol{\alpha}$ for $z > 0$. Then, it is easy to see that

$$\boldsymbol{p}(z) = z^{-1}\boldsymbol{p}(z)\boldsymbol{t}\boldsymbol{\alpha}(z^{-1}\theta(z)I - T)^{-1}, \qquad \boldsymbol{q}(z) = z^{-1}\boldsymbol{\alpha}\boldsymbol{q}(z)(z^{-1}\theta(z)I - T)^{-1}\boldsymbol{t}. \quad (4.16)$$

Note that $\boldsymbol{p}(z)\boldsymbol{t}$ and $\boldsymbol{\alpha}\boldsymbol{q}(z)$ are scalars. These expressions of the eigenvectors are well known (see, e.g., Section 5 of [19]). □

A typical queueing model with phase type renewal arrivals and services is $PH/PH/c$ queue, where PH stands for the phase type renewal process (see, e.g., [19]). If we truncate the buffer of this queue by K, then we have a $PH/PH/c/K+c$ queue, which is of course a special case of the $MArP/MSP/c/K+c$ queue. So, Theorem 4.1 and Corollary 4.1 are valid for them.

In the rest of this section, we compare our results with Theorem 1 of [4], which gives asymptotic behaviors for a $MArP/G/1/K+1$ queue. If the service time distribution is limited to phase type, then this model becomes a special case of $MArP/MSP/c/K+c$. Thus, Theorem 4.1 and Corollary 4.1 can be considered as a generalization of Theorem 1 of [4]. Here are two generalizations, one from a single server to many server, and another from independent and identically distributed service times to a Markovian service process. However, it is not immediate to conclude Theorem 1 of [4] from Theorem 4.1 or Corollary 4.1 since the modeling is different.

So far, we show how Theorem 1 of [4] for phase type service times is derived from Theorem 4.1. To this end, we first introduce $MArP/PH/1/K+1$ formally.

Example 4.2 ($MArP/PH/1/K+1$ queue) For the $MArP/MSP/c/K+c$ queue with $c=1$, rewrite C_2 as T_2, and let $D_2 = \boldsymbol{t}_2\boldsymbol{\alpha}_2$ for a probability vector $\boldsymbol{\alpha}_2$ on $\mathcal{M}_2$ and $\boldsymbol{t}_2 = -T_2\mathbf{1}$. It is easy to see that $\boldsymbol{\nu}_2 = \mu\boldsymbol{\alpha}_2(-T_2)^{-1}$ for $i=1,2$, where $\mu = (\boldsymbol{\alpha}_2(-T_2)^{-1}\mathbf{1})^{-1}$. Let

$$\underline{Q}_{-1} = I_1 \otimes \boldsymbol{t}_2, \qquad \underline{Q}_0 = C_1, \qquad \underline{Q}_1 = D_1 \otimes \boldsymbol{\alpha}_2.$$

That is, the service time of a customer who found the system empty is independently sampled from the distribution $PH(\boldsymbol{\alpha}_2, T_2)$. We refer to this model as the $MArP/PH/1/K+1$ queue. □

We apply Theorem 4.1 to the $MArP/PH/1/K+1$ queue. Similarly to (4.16), we have, from (4.4),

$$z^{-1}\boldsymbol{q}_2(z) = \left(z^{-1}\theta_2(z)I_2 - T_2\right)^{-1}\boldsymbol{t}_2\boldsymbol{\alpha}_2\boldsymbol{q}_2(z). \tag{4.17}$$

Let $s_i(z) = z^{-1}\theta_i(z)$ for $i=1,2$. Let V denote a generic random variable subject to the service time distribution $PH(\boldsymbol{\alpha}_2, T_2)$. Multiplying $\boldsymbol{\alpha}_2$ to (4.17), and noting that $\boldsymbol{\alpha}_2\boldsymbol{q}_2 > 0$, we have

$$z^{-1} = E(e^{-s_2(z)V}).$$

Differentiating both sides of this equation at $z=\eta$ yields

$$\theta_2'(\eta) = -s_1(\eta) + \frac{1}{\eta E(Ve^{s_1(\eta)V})}, \tag{4.18}$$

since $s_1(\eta) + s_2(\eta) = 0$. Let $\eta \in (0,1)$ be the solution of (4.7), and $\delta_2 = \eta\boldsymbol{\alpha}_2\boldsymbol{q}_2(\eta)$. Then, from (4.17) with $z=\eta$, we have

$$\boldsymbol{q}_2(\eta) = \delta_2\left(s_2(\eta)I_2 - T_2\right)^{-1}\boldsymbol{t}_2. \tag{4.19}$$

Similarly, letting $\gamma_2 = \eta\boldsymbol{p}_2(\eta)\boldsymbol{t}_2$, (4.3) yields

$$\boldsymbol{p}_2(\eta) = \gamma_2\boldsymbol{\alpha}_2(s_2(\eta)I_2 - T_2)^{-1}.$$

We normalize $\boldsymbol{p}_2(\eta)$ and $\boldsymbol{q}_2(\eta)$ by (4.13), then it can be obtained that

$$\gamma_2 = \frac{\theta_1(\eta)}{1-\eta}, \qquad \delta_2 = \frac{1-\eta}{\theta_1(\eta)E(Ve^{s_1(\eta)V})}. \tag{4.20}$$

The derivations of these formulas are given in Appendix D. Define

$$\chi(z) = E(e^{z\theta_1(1/z)V}), \qquad z > 0.$$

This $\chi(z)$ is the Perron-Frobenius eigenvalue of the matrix moment generating function:

$$A(z) \equiv E(e^{(C_1+zD_2)V}).$$

See [4] for details. It is straightforward to compute the derivative of $\chi(z)$ at $z = 1/\eta$ as

$$\begin{aligned}\chi'\left(\frac{1}{\eta}\right) - 1 &= (\theta_1(\eta) - \eta\theta_1'(\eta))E(Ve^{s_1(\eta)V}) - 1 \\ &= \eta(s_1(\eta) - \theta_1'(\eta))E(Ve^{s_1(\eta)V}) - 1 \\ &= \eta E(Ve^{s_1(\eta)V})\left(s_1(\eta) - \frac{1}{\eta E(Ve^{s_1(\eta)V})} - \theta_1'(\eta)\right) \\ &= -\eta E(Ve^{s_1(\eta)V})(\theta_1'(\eta) + \theta_2'(\eta)), \end{aligned} \tag{4.21}$$

where the last equality is obtained by (4.18). We are now ready to present the following result, which is proved in Appendix E.

Corollary 4.2 For $MArP/PH/1/K+1$ with the normalization (4.13), if $\rho < 1$, then we have

$$\lim_{K\to\infty} \eta^{-(K+1)} p_{\text{loss}}^{(K+1)} = \frac{(1-\rho)s_1(\eta)}{\lambda(1-\eta)\left(\chi'\left(\frac{1}{\eta}\right) - 1\right)} \boldsymbol{\pi}_0 \boldsymbol{q}_1(\eta)\boldsymbol{p}_1(\eta)\hat{\boldsymbol{r}}_1, \tag{4.22}$$

where $\hat{\boldsymbol{r}}_1$ is the k_1-dimensional column vector such that $\hat{\boldsymbol{r}} = \hat{\boldsymbol{r}}_1 \otimes \hat{\boldsymbol{r}}_2$ and $\boldsymbol{\nu}_1\hat{\boldsymbol{r}}_1 = 1$.

This corollary must be identical with Theorem 1 of [4]. However, it is still apart from (4.22). We here need to make correspondences between their and our models. In [4], the queue is described by the embedded Markov chain at departure epochs. This embedded Markov chain is a truncation of the so called $M/G/1$ type queue. In what follows, we show that they are indeed identical, introducing some notation of the $M/G/1$ type queue,

Let $\{(L_n, J_n)\}$ be a Markov chain for the corresponding $M/G/1$ type queue with an unlimited buffer, where L_n and J_n represent the number of customers just after departure and the phase for arrivals, respectively. If L_n remains in the set $\{1, 2, \ldots\}$, then $\{(L_n, J_n)\}$ can be considered as a Markov additive process. Denote this additive process by (X_n, B_n). For this additive process, we can define matrix G_{D} corresponding to the G for hitting states in one lower level. Similarly, we can define matrix $\hat{R}_{\text{D}}$ corresponding to the $\hat{R}$ for occupation measures. Since this unlimited buffer queue is stable, it is not hard to see that G_{D} is a proper transition probability matrix, and $\boldsymbol{\nu}_1$ is the left-invariant vector of $\hat{R}_{\text{D}}$.

Let $\boldsymbol{g}_{\text{D}}$ and $\hat{\boldsymbol{r}}_{\text{D}}$ be the left-invariant and right-invariant vectors of G_{D} and $\hat{R}_{\text{D}}$, respectively, such that $\boldsymbol{g}_{\text{D}}\mathbf{1} = 1$ and $\boldsymbol{\nu}\hat{\boldsymbol{r}}_{\text{D}} = 1$. Then, as is noticed in [4], it is well known that

$$\boldsymbol{\pi}_0 = (1-\rho)\boldsymbol{g}_{\text{D}}. \tag{4.23}$$

In Appendix F, we show that

$$\hat{\boldsymbol{r}}_1 = \hat{\boldsymbol{r}}_{\text{D}}. \tag{4.24}$$

Substituting these formulas into (4.22) and using $s_1(\eta) = \eta^{-1}\theta_1(\eta)$, we get

$$\lim_{K\to\infty} \eta^{-K} p_{\text{loss}}^{(K+1)} = \frac{(1-\rho)^2\theta_1(\eta)}{\lambda(1-\eta)\left(\chi'\left(\frac{1}{\eta}\right) - 1\right)} \boldsymbol{g}_{\text{D}}\boldsymbol{q}_1(\eta)\boldsymbol{p}_1(\eta)\hat{\boldsymbol{r}}_{\text{D}}.$$

This is exactly the result in Theorem 1 of [4]. It may be notable that the loss probability asymptotics have different expressions. Among them, (4.10) may be most tractable for numerical computations. It also has an advantage to be valid for many server queues.

5. Concluding remarks

We finally consider possible extensions of the present results. The QBD process with finitely many background states is sufficient for many queueing applications, but it is still limited since it excludes batch arrivals and batch departures. To include them, we may need unbounded level jumps. This class of queueing models are called $GI/G/1$ type (see, e.g., [14]). An extension of the present results for such models may be interesting. Another extension is to accommodate continuous quantities in background states. For example, they are required for the models with general service time distributions. For such an extension, we need to enlarge the framework of the Markov additive process of this paper. This is also a challenging subject.

Appendix A. Proof of Proposition 3.1

Dividing $Q^{(m)}$ into level 0, $\{1,\dots,m-1\}$ and m, we have

$$\begin{pmatrix} \underline{Q}_0 & \underline{Q}_1 & & & \\ \underline{Q}_{-1} & & & & \\ & & Q_*^{(m)} & & \\ & & & & Q_1 \\ & & & \overline{Q}_{-1} & \overline{Q}_0 \end{pmatrix}, \tag{A.1}$$

where $Q_*^{(m)}$ consists of $(m-1)$ blocks in each rows and columns such that

$$Q_*^{(m)} = \begin{pmatrix} Q_0 & Q_1 & & & \\ Q_{-1} & Q_0 & Q_1 & & \\ & \ddots & \ddots & \ddots & \\ & & Q_{-1} & Q_0 & Q_1 \\ & & & Q_{-1} & Q_0 \end{pmatrix}.$$

Clearly, the stationary distribution $\boldsymbol{\pi}^{(m)}$ satisfies $\boldsymbol{\pi}^{(m)}Q^{(m)} = \mathbf{0}$, which is equivalent to

$$\boldsymbol{\pi}_0^{(m)}\underline{Q}_0 + \boldsymbol{\pi}_1^{(m)}\underline{Q}_{-1} = \mathbf{0}, \quad \boldsymbol{\pi}_{m-1}^{(m)}Q_1 + \boldsymbol{\pi}_m^{(m)}\overline{Q}_0 = \mathbf{0}, \tag{A.2}$$

$$(\boldsymbol{\pi}_0^{(m)}\underline{Q}_1, \mathbf{0}, \dots, \mathbf{0}) + (\boldsymbol{\pi}_1^{(m)}, \dots, \boldsymbol{\pi}_{m-1}^{(m)})Q_*^{(m)} + (\mathbf{0}, \dots, \mathbf{0}, \boldsymbol{\pi}_m^{(m)}\overline{Q}_{-1}) = \mathbf{0}. \tag{A.3}$$

Since $Q_*^{(m)}$ is a defective matrix, $(-Q_*^{(m)})^{-1}$ exists. Thus, (A.3) becomes

$$(\boldsymbol{\pi}_1^{(m)}, \dots, \boldsymbol{\pi}_{m-1}^{(m)}) = (\boldsymbol{\pi}_0^{(m)}\underline{Q}_1, \mathbf{0}, \dots, \mathbf{0}, \boldsymbol{\pi}_m^{(m)}\overline{Q}_{-1})(-Q_*^{(m)})^{-1}. \tag{A.4}$$

Note that $(-Q_*^{(m)})^{-1}$ represents the mean sojourn time in $(\mathcal{N}_m \setminus \{0,m\}) \times \mathcal{S}_B$ before absorbing to $\{0\} \times \mathcal{S}_B$ or $\{m\} \times \mathcal{S}_B$. So, (A.2) and (A.4) prove this proposition.

Appendix B. Proof of (3.10)

We formally prove (3.11):

$$N_{1m}^{(0,+)} = N_{1(m-1)}^{(0,m)} Q_1 N_{mm}^{(0,+)}, \qquad m \geq 2.$$

For convenience, we let $\tau_0^- = \tau^c_{[1,\infty)}$ and $\tau_\ell^+ = \tau^c_{(-\infty,\ell-1]}$ for $\ell \geq 1$. That is,

$$\tau_0^- = \inf\{t > 0; X(t) \leq 0\}, \qquad \tau_\ell^+ = \inf\{t > 0; X(t) \geq \ell\}.$$

Then, conditioning the Markov additive process $(X(t), B(t))$ at time τ_m^+, we have

$$\begin{aligned}
N_{1m}^{(0,+)}(i,j) &= E_{(1,i)}\left(\int_0^\infty 1(X(u) = m, B(u) = j, u < \tau_0^-)du\right) \\
&= \int_0^\infty E_{(1,i)}\left(P_{(X(\tau_m^+),B(\tau_m^+))}(X(u) = m, B(u) = j, u < \tau_0^-)\right) du \\
&= E_{(1,i)}\left(\int_{\tau_m^+}^\infty P_{(X(\tau_m^+),B(\tau_m^+))}(X(u) = m, B(u) = j, u < \tau_0^-)du 1(\tau_m^+ < \tau_0^-)\right) \\
&= \sum_k P_{(1,i)}\left(B(\tau_m^+) = k, \tau_m^+ < \tau_0^-\right) N_{mm}^{(0,+)}(k,j).
\end{aligned}$$

Hence, (3.11) is obtained if we show that

$$P_{(1,i)}\left(B(\tau_m^+) = k, \tau_m^+ < \tau_0^-\right) = N_{1(m-1)}^{(0,m)} Q_1(i,k). \tag{B.1}$$

Let A be the transition rate matrix of $(X(t), B(t))$. It is easy to see that

$$A((k,i),(\ell,j)) = \begin{cases} Q_{-1}(i,j), & \ell = k-1, \\ Q_1(i,j), & \ell = k, \\ Q_{+1}(i,j), & \ell = k+1, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\sigma_t = \tau_0^- \wedge \tau_m^+ \wedge t$ for $t > 0$, where $a \wedge b = \min(a,b)$ for real number a, b. Since σ_t is a bounded stopping time with respect to the filtration of the Markov chain $(X(t), B(t))$, we have the following Dynkin's formula (e.g., see Lemma 19.21 of [9]).

$$E_{(\ell,i)}(f(X(\sigma_t), B(\sigma_t)) = f(\ell,i) + E_{(\ell,i)}\left(\int_0^{\sigma_t} Af(X(u), B(u))du\right), \tag{B.2}$$

where f is a nonnegative and bounded function on $\mathcal{N} \times \mathcal{S}_B$, and

$$Af(\ell,i) = \sum_{(\ell',i') \in \mathcal{N}\times\mathcal{S}_B} A((\ell,i),(\ell',i'))f(\ell',i').$$

Let $\ell = 1$ and $f(\ell',i') = 1(\ell' = m, i' = j)$ in (B.2), then we have, for $m \geq 2$,

$$\begin{aligned}
&P_{(1,i)}(X(\sigma_t) = m, B(\sigma_t) = k) \\
&\quad = E_{(1,i)}\left(\int_0^{\sigma_t} A((X(u), B(u)), (m,k))du\right) \\
&\quad = \sum_{i'} \int_0^\infty P_{(1,i)}\left(X(u) = m-1, B(u) = i', u < \sigma_t\right) Q((m-1,i'),(m,k))du.
\end{aligned}$$

Letting $t \to \infty$ in the above formula and using the fact that $X(\tau_0^-) = 0 \neq m$, we have

$$\begin{aligned} &P_{(1,i)}(B(\tau_m^+) = k, \tau_m^+ < \tau_0^-) \\ &\quad = \sum_{i'} E_{(1,i)} \left(\int_0^\infty 1(X(u) = m-1, B(u) = i, u < \tau_m^+ \wedge \tau_0^-))du \right) \\ &\qquad\qquad \times Q((m-1, i'), (m, k)) \\ &\quad = N^{(0,m)}_{1(m-1)} Q_1(i, k). \end{aligned}$$

This completes the proof of (3.11).

Appendix C. Proof of (3.14)

We show that (3.14) agrees with (3.4) for the $M/PH/1/m$ queue. This is equivalent to show that

$$\eta \boldsymbol{p}(\eta^{-1} I - \hat{R}) Q_1 = \lambda \boldsymbol{p} T^{-1}(Q_0 + Q_1 + \hat{R} Q_1)$$

We here prefer (3.4) to (3.12). Since $Q_0 + Q_1 = T$ and $Q_1 = \lambda I$, this is further equivalent to

$$\boldsymbol{p}(\eta I + \lambda T^{-1}) \hat{R} Q_1 = \boldsymbol{0}.$$

Hence, it is sufficient to show that

$$\boldsymbol{p}(\eta I + \lambda T^{-1}) T \boldsymbol{1} \boldsymbol{\alpha} \hat{G} = \boldsymbol{0}. \tag{C.1}$$

since $\hat{R} Q_1 = Q_{-1} \hat{G}$ and $Q_{-1} = \boldsymbol{t}\boldsymbol{\alpha}$. We show that $\boldsymbol{p}(\eta I + \lambda T^{-1}) T \boldsymbol{1} = 0$. To this end, we note that $\boldsymbol{p}$ is determined by (2.5), which is written as

$$\boldsymbol{p} \left(\lambda(1-\eta) I + \eta T + \eta^2 \boldsymbol{t}\boldsymbol{\alpha} \right) = \boldsymbol{0}.$$

We normalize $\boldsymbol{p}$ in such a way that $\boldsymbol{p}\boldsymbol{t} = \eta^{-1}$. Then, postmultiplying both sides of this with $\boldsymbol{1}$, we have

$$\boldsymbol{p}(\lambda I + \eta T) \boldsymbol{1} = \eta(\lambda \boldsymbol{p} \boldsymbol{1} - 1).$$

This yields

$$(1 - \eta)(\lambda \boldsymbol{p} \boldsymbol{1} - 1) = 0,$$

which implies that $\lambda \boldsymbol{p} \boldsymbol{1} = 1$. Hence, we have $\boldsymbol{p}(\eta T + \lambda I) \boldsymbol{1} = 0$, so we have (C.1).

Appendix D. Proof of (4.20)

From the normalization, we have

$$\boldsymbol{\alpha}_2 (s_2(\eta) I_2 - T_2)^{-1} \boldsymbol{1}_2 = \gamma_2^{-1},$$

while (4.17) yields

$$\boldsymbol{\alpha}_2(s_2(\eta)I_2 - T_2)^{-1}(-T_2)\mathbf{1}_2 = \eta^{-1}.$$

These two equations imply

$$\begin{aligned} 1 &= \boldsymbol{\alpha}_2\mathbf{1}_2 \\ &= \boldsymbol{\alpha}_2(s_2(\eta)I_2 - T_2)^{-1}(s_2(\eta)I_2 - T_2)\mathbf{1}_2 \\ &= s_2(\eta)\gamma_2^{-1} + \eta^{-1}. \end{aligned}$$

This concludes the first equation of (4.20) since $\eta s_2(\eta) = \theta_2(\eta)$ and $\theta_1(\eta) = -\theta_2(\eta)$. We next compute the inner product of $\boldsymbol{p}_2(\eta)$ and $\boldsymbol{q}_2(\eta)$.

$$\begin{aligned} \boldsymbol{p}_2(\eta)\boldsymbol{q}_2(\eta) &= \gamma_2\delta_2\boldsymbol{\alpha}_2(s_2(\eta)I_2 - T_2)^{-2}\boldsymbol{t}_2 \\ &= -\gamma_2\delta_2\frac{d}{dz}\boldsymbol{\alpha}_2(zI_2 - T_2)^{-1}\boldsymbol{t}_2\Big|_{z=s_2(\eta)} \\ &= -\gamma_2\delta_2\frac{d}{ds}E(e^{-sB})\Big|_{s=s_2(\eta)} \\ &= \gamma_2\delta_2 E(Be^{s_1(\eta)B}). \end{aligned}$$

Hence, the normalization $\boldsymbol{p}_2(\eta)\boldsymbol{q}_2(\eta) = 1$ concludes the second equation of (4.20).

Appendix E. Proof of Corollary 4.2

From Corollary 4.1 and (4.21), we have

$$\lim_{K\to\infty}\eta^{-(K+1)}p_{\text{loss}}^{(K+1)} = \frac{\eta(1-\rho)E(Be^{s_1(\eta)B})\boldsymbol{\pi}_1\boldsymbol{r}\boldsymbol{p}\hat{\boldsymbol{r}}}{\rho(1-\eta)\left(\chi'\left(\frac{1}{\eta}\right) - 1\right)}.$$

Hence, if we show that

$$\boldsymbol{\pi}_1\boldsymbol{r} = \frac{(1-\eta)}{\eta E(Be^{s_1(\eta)B})}\boldsymbol{\pi}_0\boldsymbol{q}_1(\eta), \tag{E.1}$$

$$\boldsymbol{p}\hat{\boldsymbol{r}} = \frac{s_1(\eta)}{\mu(1-\eta)}\boldsymbol{p}_1(\eta)\hat{\boldsymbol{r}}_1, \tag{E.2}$$

then Corollary 4.2 is obtained. From (2.11), we have

$$\boldsymbol{\pi}_1 = \boldsymbol{\pi}_0(I_1 \otimes \boldsymbol{\alpha}_2)Q_1N_{11}^{(0,+)} = \boldsymbol{\pi}_0(I_1 \otimes \boldsymbol{\alpha}_2)R.$$

Since $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0 \otimes 1$, this, (2.7) and (3.5) yield

$$\begin{aligned} \boldsymbol{\pi}_1\boldsymbol{r} &= \eta(\boldsymbol{\pi}_0 \otimes \boldsymbol{\alpha}_2)\boldsymbol{r} \\ &= -\eta(\boldsymbol{\pi}_0 \otimes \boldsymbol{\alpha}_2)(Q_0 + \eta Q_{-1} + Q_1G)(\boldsymbol{q}_1(\eta) \otimes \boldsymbol{q}_2(\eta)) \\ &= -\eta(\boldsymbol{\pi}_0 \otimes \boldsymbol{\alpha}_2)(C_1 \oplus T_2 + \eta I_1 \otimes \boldsymbol{t}_2\boldsymbol{\alpha}_2 + (D_1 \otimes I_2)G)(\boldsymbol{q}_1(\eta) \otimes \boldsymbol{q}_2(\eta)). \end{aligned} \tag{E.3}$$

Since G is the transition probabilities to hit one lower level, which always ends up with the departure of a customer, G can be decomposed as

$$G = G_1 \otimes \mathbf{1}_2 \boldsymbol{\alpha}_2$$

for some $k_1 \times k_1$ transition probability matrix G_1. Applying this to (2.10), we have

$$\begin{aligned}
\mathbf{0} &= \boldsymbol{\pi}_0(C_1 + (I_1 \otimes \mathbf{1}_1\boldsymbol{\alpha}_2)R(I_1 \otimes \boldsymbol{t}_2)) \\
&= \boldsymbol{\pi}_0(C_1 + (I_1 \otimes \mathbf{1}_1\boldsymbol{\alpha}_2)RQ_{-1}(I_1 \otimes \mathbf{1}_2)) \\
&= \boldsymbol{\pi}_0(C_1 + (I_1 \otimes \mathbf{1}_1\boldsymbol{\alpha}_2)Q_1G(I_1 \otimes \mathbf{1}_2)) \\
&= \boldsymbol{\pi}_0(C_1 + (D_1 \otimes \mathbf{1}_1\boldsymbol{\alpha}_2)(G_1 \otimes \mathbf{1}_2) \\
&= \boldsymbol{\pi}_0(C_1 + D_1G_1).
\end{aligned}$$

Hence, we have $\boldsymbol{\pi}_0 C_1 = -\boldsymbol{\pi}_0 D_1 G_1$. Substituting this to (E.3) together with (4.19) and $\delta_2 = \eta\boldsymbol{\alpha}_2\boldsymbol{q}_2(\eta)$, we get

$$\begin{aligned}
\boldsymbol{\pi}_1\boldsymbol{r} &= -\eta(\boldsymbol{\pi}_0 \otimes \boldsymbol{\alpha}_2)(I_1 \otimes T_2 + \eta I_1 \otimes \boldsymbol{t}_2\boldsymbol{\alpha}_2)(\boldsymbol{q}_1(\eta) \otimes \boldsymbol{q}_2(\eta)) \\
&= -\eta(\boldsymbol{\pi}_0 \otimes \boldsymbol{\alpha}_2)(I_1 \otimes T_2 + I_1 \otimes (\eta\boldsymbol{t}_2\boldsymbol{\alpha}_2))(\boldsymbol{q}_1(\eta) \otimes \boldsymbol{q}_2(\eta)) \\
&= -\eta\boldsymbol{\pi}_0\boldsymbol{q}_1(\eta)\boldsymbol{\alpha}_2(T_2 + \eta\boldsymbol{t}_2\boldsymbol{\alpha}_2)\boldsymbol{q}_2(\eta) \\
&= -\eta\boldsymbol{\pi}_0\boldsymbol{q}_1(\eta)(\boldsymbol{\alpha}_2 T_2\boldsymbol{q}_2(\eta) + \delta_2\boldsymbol{\alpha}_2\boldsymbol{t}_2).
\end{aligned}$$

We further compute this using the fact that

$$\begin{aligned}
\boldsymbol{\alpha}_2 T_2\boldsymbol{q}_2(\eta) &= \delta_2\boldsymbol{\alpha}_2 T_2 \left(s_2(\eta)I_2 - T_2\right)^{-1}\boldsymbol{t}_2 \\
&= -\delta_2\boldsymbol{\alpha}_2 \left(s_2(\eta)I_2 - T_2\right)\left(s_2(\eta)I_2 - T_2\right)^{-1}\boldsymbol{t}_2 + \delta_2 s_2(\eta)\boldsymbol{\alpha}_2 \left(s_2(\eta)I_2 - T_2\right)^{-1}\boldsymbol{t}_2 \\
&= -\delta_2\boldsymbol{\alpha}_2\boldsymbol{t}_2 + \delta_2 s_2(\eta)\boldsymbol{\alpha}_2 \left(s_2(\eta)I_2 - T_2\right)^{-1}\boldsymbol{t}_2 \\
&= -\delta_2\boldsymbol{\alpha}_2\boldsymbol{t}_2 - \delta_2 s_1(\eta)\eta^{-1},
\end{aligned}$$

we get

$$\boldsymbol{\pi}_1\boldsymbol{r} = \delta_2 s_1(\eta)\boldsymbol{\pi}_0\boldsymbol{q}_1(\eta).$$

Hence, (4.20) concludes (E.1).

We next prove (E.2). To this end, we shall use dual processes for the Markov additive processes $(X(t), B(t))$ and $(\hat{X}(t), B(t))$. These processes are defined as $(-X(-t), B(-t))$ and $(-\hat{X}(-t), B(-t))$, respectively. Obviously, $(-X(-t), B(-t))$ is also a Markov additive process generated by matrices:

$$Q_i^* = \Delta_{\boldsymbol{\nu}}^{-1}(Q_i)^{\mathrm{T}}\Delta_{\boldsymbol{\nu}}, \qquad i = 0, \pm 1,$$

where A^{T} is the transpose of matrix A, and $\Delta_{\boldsymbol{a}}$ is the diagonal matrix whose i-th diagonal entry is the i-th entry of a vector $\boldsymbol{a}$. We use superscript "$*$" to indicate the dual process. Similarly, $(-\hat{X}(-t), B(-t))$ is generated by matrices:

$$\hat{Q}_i^* = \Delta_{\boldsymbol{\nu}}^{-1}(Q_{-i})^{\mathrm{T}}\Delta_{\boldsymbol{\nu}}, \qquad i = 0, \pm 1.$$

Let $\hat{N}_{00}^{*(-,1)}$ be the corresponding matrix with $N_{00}^{(-,1)}$ for the dual direction reversed process $(-\hat{X}(-t), B(-t))$. Define

$$\hat{G}_+^* = \hat{N}_{00}^{*(-,1)}\hat{Q}_1^*.$$

This matrix represents the background state transition probabilities at hitting one upper level for $(-\hat{X}(-t), B(-t))$. On the other hand, it can be shown that

$$\hat{R} = \Delta_{\boldsymbol{\nu}}^{-1}(\hat{G}_+^*)^{\mathrm{T}}\Delta_{\boldsymbol{\nu}}.$$

For example, see [14] for a discrete time version of this relation, which is easily converted to continuous time versions. Since $\hat{G}_+^* = G^*$ can be obtained by reversing the direction, this becomes

$$\hat{R} = \Delta_{\boldsymbol{\nu}}^{-1}(G^*)^{\mathrm{T}}\Delta_{\boldsymbol{\nu}}. \tag{E.4}$$

We now closely look at G^*. We first observe that G^* is a non-defective transition probability matrix since the dual process has the same mean drift as the original process. We next note that

$$\begin{aligned} Q_{-1}^* &= \Delta_{\boldsymbol{\nu}}^{-1}(I_1 \otimes \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{t}_2^{\mathrm{T}})\Delta_{\boldsymbol{\nu}} \\ &= I_1 \otimes \Delta_{\boldsymbol{\nu}_2}^{-1}\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{t}_2^{\mathrm{T}}\Delta_{\boldsymbol{\nu}_2}, \end{aligned}$$

where $\boldsymbol{\nu} = \boldsymbol{\nu}_1 \otimes \boldsymbol{\nu}_2$ is used. Since

$$\boldsymbol{t}_2^{\mathrm{T}}\Delta_{\boldsymbol{\nu}_2}\mathbf{1} = \boldsymbol{\nu}_2\boldsymbol{t}_2 = \mu,$$

G^* must has the following form for some $k_1 \times k_1$ transition probability matrix G_1^*.

$$G^* = G_1^* \otimes \mathbf{1}_2\boldsymbol{\xi}_2,$$

where $\boldsymbol{\xi}_2 = \mu^{-1}\boldsymbol{t}_2^{\mathrm{T}}\Delta_{\boldsymbol{\nu}_2}$. Let $\boldsymbol{g}^*$ and $\boldsymbol{g}_1^*$ be the left invariant vectors of G^* and G_1^*, respectively, then the left invariant vector $\boldsymbol{g}^*$ of G^* is given by

$$\boldsymbol{g}^* = \boldsymbol{g}_1^* \otimes \boldsymbol{\xi}_2. \tag{E.5}$$

Hence, (E.4) implies that the right invariant vector $\hat{r}$ of $\hat{R}$ is given by

$$\hat{\boldsymbol{r}} = a\Delta_{\boldsymbol{\nu}}^{-1}(\boldsymbol{g}^*)^{\mathrm{T}} = a\hat{\boldsymbol{r}}_1 \otimes (\mu^{-1}\boldsymbol{t}_2),$$

for some positive constant a, where $\hat{\boldsymbol{r}}_1 = \Delta_{\boldsymbol{\nu}_1}^{-1}(\boldsymbol{g}_1^*)^{\mathrm{T}}$. Since $\boldsymbol{\nu}\hat{\boldsymbol{r}} = 1$, we have $a = 1$. These yield

$$\begin{aligned} \boldsymbol{p}\hat{\boldsymbol{r}} &= \mu^{-1}\boldsymbol{p}_1(\eta)\hat{\boldsymbol{r}}_1\boldsymbol{p}_2(\eta)\boldsymbol{t}_2 \\ &= \mu^{-1}\gamma_2\eta^{-1}\boldsymbol{p}_1(\eta)\hat{\boldsymbol{r}}_1. \end{aligned}$$

Thus, (4.20) concludes (E.2).

Appendix F. Proof of (4.24)

From (E.4) and (E.5), we have $\hat{\boldsymbol{r}}_1 = \Delta_{\boldsymbol{\nu}_1}^{-1}\boldsymbol{g}_1^*$. Let $\boldsymbol{g}_{\mathrm{D}}^*$ be the corresponding vector with $\boldsymbol{g}_{\mathrm{D}}$ for the dual of the Markov additive process (X_n, B_n). It can be shown that the background processes for these additive processes have the common stationary distribution $\boldsymbol{\nu}_1$. So, we have $\hat{\boldsymbol{r}}_{\mathrm{D}} = \Delta_{\boldsymbol{\nu}_1}^{-1}\boldsymbol{g}_{\mathrm{D}}^*$ similarly to $\hat{\boldsymbol{r}}_1$ (see also [4]). Since both of $(1-\rho)\boldsymbol{g}^*$ and $(1-\rho)\boldsymbol{g}_{\mathrm{D}}^*$ give the same stationary probabilities of background states in the dual process when the system is empty, they are identical, and we have (4.24).

Acknowledgements This research is supported in part by JSPS under grant No. 13680532.

References

[1] Asmussen, S. *Applied Probability of Queues*, Springer, New York, 2003, 2nd Edition.

[2] Bellman, R. *Introduction to Matrix Analysis*, SIAM, Philadelphia, 1997, 2nd Edition.

[3] Baiocchi, A. Asymptotic behaviour of the loss probability of the $M/G/1/K$ and $G/M/1/K$ queues, *Queueing Systems* **1992**, 10, 235-248.

[4] Baiocchi, A. Analysis of the loss probability of the $MAP/G/1/K$ queues, Part I: Asymptotic theory, *Stochastic Models* **1994**, 10, 867-893.

[5] Baiocchi, A. and Blefari-Melazzi, N. Analysis of the loss probability of the $MAP/G/1/K$ queues, Part II: Approximations and numerical results, *Stochastic Models* **1994**, 10, 895-925.

[6] Choi, B.D., Kim, B. and Wee, I. Asymptotic behavior of loss probability in $GI/M/1/K$ queue as k tends to infinity, *Queueing Systems* **2000**, 36, 437-442.

[7] Çinlar, E. (1972) Markov Additive processes II, *Z. Wahrscheinlichkeitstheorie* **1972**, 24, 437-442.

[8] Hajek, B. Birth-and death processes on the integers with phases and general boundaries, *Journal of Applied Probability* **1982**, 19, 488–499.

[9] Kallenberg, O. *Foundation of Modern Probability*, Springer, New York, 2001, 2nd Edition.

[10] Kingman, J.F.C. A convexity property of positive matrices, *Quarterly journal of mathematics* **1961**, 12, 283-284.

[11] Latouche, G. and Ramaswami, V. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, American Statistical Association and the Society for Industrial and Applied Mathematics, Philadelphia, **1999**.

[12] Li, H. and Zhao, Y.Q. Stochastic block-monotonicity in the approximation of the stationary distribution of infinite Markov chains, *Stochastic Models* **2000**, 16, 313–333.

[13] Miyazawa, M. and Tijms, H. Comparison of two approximations for the loss probability in finite-buffer queues. *Prob. Eng. Inf. Sci.* **1993**, 7, 19-27.

[14] Miyazawa, M. and Zhao, Y.Q. The stationary tail asymptotics in the $GI/G/1$ type queue with countably many background states, *Advances in Applied Probability* **2004**, 36, 1231–1251.

[15] Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models*, Johns Hopkins University Press, Baltimore, 1981.

[16] Ozawa, T. Analysis of Queues with Markovian Service Processes, *Stochastic Models* **2004**, 20, 391–423.

[17] Sakasegawa, H., Miyazawa, M. and Yamazaki, G. Evaluating the Overflow probability using the infinite queue. *Management Science* **1993**, 39, 1238-1245.

[18] Takács, L. *Introduction to the Theory of Queues*, Oxford University Press, New York, 1962.

[19] Takahashi, Y. Asymptotic exponentiality of the tail of the waiting-time distribution in a $PH/PH/c$ queue. *Advances in Applied Probability* **1981**, 13, 619–630.

[20] Takahashi, Y., Fujimoto, K. and Makimoto, N. Geometric decay of the steady-state probabilities in a quasi-birth-and-death process with a countable number of phases, *Stochastic Models* **2001**, 17, 1–24.

[21] Wolf, D. Approximation of the invariant probability measure of an infinite stochastic matrix. *Advances in Applied Probability* **1980**, 12, 710-726.